Berry J, Budgen D, Holliman N.

Evaluating subjective impressions of quality controlled 3D films on large and small screens.

***Journal of Display Technology* 2015, 11(11), 927-938**

DOI link to article:

http://dx.doi.org/10.1109/JDT.2014.2384531

Date deposited:

31/03/2016

Evaluating subjective impressions of quality controlled 3D films on large and small screens.

Jonathan Berry, David Budgen, *Member, IEEE Computer Society* and
Nick Holliman, *Member, IEEE Computer Society.*

***Abstract*—We investigate audiences' subjective impressions of two high quality 3D stereoscopic films when viewed on a large projected display (160" diagonal). We further show that our results are repeatable using TV sized displays (50" diagonal) and sites outside our laboratory.The literature proposes a number of algorithms that aim to produce high quality stereoscopic depth in 3D films. Many of these algorithms limit the stereoscopic depth to a defined depth budget, which can be dynamically allocated through the course of a film. However, there have been no detailed studies evaluating audiences' subjective impressions of 3D films that utilise such algorithms - something we seek to correct in this study.This study comprises of an original experiment and four differentiated replications, across which we vary the film, display technology and international location used. All of these experiments implement a pre-test post-test quasi-experiment design, in which participants were asked, before and after viewing a 3D film, to rate their agreement with five statements concerning 3D films. These statements addressed the viewing experience, comfort, naturalness, suitability to conveying complex information and benefit to learning associated with 3D displays. One of two possible films were shown to each participant, both of which were produced using our own stereoscopic depth control algorithms and have won national or international awards, giving independent confirmation of their quality. Our results indicate that audiences' responses to our five statements change positively after viewing high quality 3D films. Furthermore, these results are repeatable for large and TV sized displays, as well as for locations outside our laboratory.We conclude that it is important to produce high quality content with a carefully controlled depth budget in order to evoke positive reactions in audiences to 3D films.**

***Index Terms*—Human Factors, Three-dimensional displays, Stereo vision, Large-screen displays, TV displays**

I. Introduction

In this study we have used a pre-test post-test quasi-experiment design to evaluate audiences' subjective impressions of stereoscopic (3D) films with "quality controlled" binocular depth, created using algorithms such as those detailed in [1] and [2]. In our experience, such films typically elicit positive responses on technical quality from both expert and non-expert audiences alike. Here, we aimed to rigorously evaluate these responses to address the following research questions:

1) Does viewing a high quality 3D film create a measurable change in audience attitudes towards 3D film?
2) Are the measured changes repeatable on displays with different sizes?
3) Can we replicate these results outside our laboratory?

We have addressed these research questions through an audience-centred study that gathers self-report responses and written comments from all audience members. Furthermore, this study incorporates an original experiment and a number of differentiated replications. As Lindsay and Ehrenberg write, replication is a crucial aspect of the scientific method that is perhaps often overlooked when evaluating subjective impressions [3]. The differentiated replications we report here, in which we vary the film, display and site used, offer insight into how generalisable our results are.

Hasenzahl [4] tells us that the study of *user experience* (and likewise *audience-experience*) is concerned with technologies that fulfil more than just instrumental needs. It is important to recognise the subjective, situated, complex and dynamic encounter that occurs between the user and the technology. As such, the user experience arises from characteristics of their internal state, the designed system and the context of interaction. Creating a good stereoscopic film viewing experience must therefore bring together the right film, display, audience and viewing environment.

For the film content, we employed two short 3D films entitled *Cosmic Cookery* and *Cosmic Origins*. These were developed by a collaboration between Physicists and Computer Scientists at Durham University, and produced using algorithms that quality control the binocular depth [5], [6]. Both films illustrate how theories of dark matter have influenced the formation and movement of stars and galaxies. They were initially created to be shown at the annual Royal Society's Summer Science Exhibition in London in 2005 and 2009 respectively, and have consistently received positive informal feedback from large, non-expert audiences. *Cosmic Cookery* won first prize in the national VizNet Visualisation Showcase 2006, whilst *Cosmic Origins* was winner of the "Best Computer Graphics Film Award" at the Stereoscopic Displays and Applications Conference 2010, San Jose, California.

For the display technology, we began by using the large 160" projected display that the films were designed to be viewed upon. Once we had used this display to establish that high quality films can have a measurable effect on audiences, we then investigated whether our results were repeatable on a 60" TV sized screen. Our displays were carefully selected for their low cross-talk and high resolution.

The audiences were made up of students and staff from the academic communities where the experimentation took place.

J. Berry and D. Budgen are with the School of Engineering and Computer Science, Durham University, United Kingdom.

N. S. Holliman is with the Department of Theatre, Film and Television, University of York, United Kingdom. (email: nick.holliman@york.ac.uk)

Manuscript received DD, MMMM, 2014; revised DD, MMMM, YYYY.

All participants were screened for stereo acuity prior to their involvement in the study. The first rounds of experimentation were undertaken in a laboratory at Durham University (UK) and then, once we had established a suitable 60" TV sized platform, we investigated whether our results were repeatable at other sites. First, we took the study to another UK site, York, and then we moved to an international location in Twente, The Netherlands. We sought to keep the environment, specifically brightness, sound volume and viewing angle, as similar as possible across all experimentation.

In the above ways, we designed an experiment that met the requirements specified by Hasenzahl [4] for content, display, audience and environment. Our report of this experiment begins with a review of related background material (section II), followed by a summary of the methodology adopted (section III), before detailing the specific setup and results of the experiment (section IV) and replications (sections V, VI and VII). We discuss the results in VIII and draw together conclusions and further avenues for research in section IX.

II. Background

A number of studies have looked into the subjective experience of viewing 3D films and we review these to identify the subjective attributes that might be most relevant to our research question: *Does viewing high quality 3D films create a measurable change in audience attitudes towards 3D film?*

A study by Seuntiens et al. [7] argues that the quality models normally used to evaluate 2D images are not sufficient for evaluating 3D images. This is because the attributes they incorporate, such as noise, blur, colour or brightness, do not relate to the added value of depth. Depth is degraded by unique stereoscopic attributes such as mismatched keystone distortion, image shear distortion or display crosstalk. They present a study proposing the use of *viewing experience* and *naturalness* as evaluative concepts in order to better reflect the added value of 3D images. A set of stereoscopic images were degraded using various amounts of additive noise and shown to participants who rated them according to viewing experience and naturalness. The ratings of viewing experience indicated significant effects existed for the amount of noise, the image shown, and whether or not the image was 3D or 2D. Naturalness yielded significant effects for the amount of noise in the image, as well as whether the image was 3D or 2D. No interactions were found between any of the effects. The study therefore concluded that both naturalness and viewing experience are useful attributes for capturing the added value of depth in 3D images.

Whilst the use of binocular cues may impact positively upon viewing experience and naturalness, they can also impact negatively upon other factors such as visual comfort, fatigue and sickness [8], [9], [10]. The film-maker Lenny Lipton writes in his book *The Foundations of Stereoscopic Cinema* that: "The danger with stereoscopic film-making is that if it is improperly done, the result can be discomfort. Yet, when properly executed, stereoscopic films are beautiful and easy on the eyes." [11] Improved visual comfort is undoubtedly a key goal for high quality 3D, and thus an important part of an audience-centred study.

The study by Polonen et. al. assessed the subjective responses of 85 participants to a 3D cinema viewing of the Hollywood blockbuster *Avatar* [12]. The participants filled out a series of questionnaires, including the Simulator Sickness Questionnaire (SSQ), before and after watching the film. The post-viewing questionnaires included questions about viewing experience, naturalness and comfort. Results from this experiment could then be compared with a similar previous experiment in which participants viewed the film *U2 3D*. It was found that viewing experience and naturalness both had average response values of approximately 7.5 out of 10. Additionally it was reported that approximately 10% of viewers may feel sick after a relatively long presentation, and that visual strain and sickness was roughly the same for the 165 minute long Avatar film and the 85 minute long U2 3D film. No reference measurements for these values were taken before the viewing, so the change in audience perceptions of 3D films before and after viewing the films was not addressed.

A group of studies investigated audience response to 3D television (3DTV), by using data collected over a three day period in a shopping mall [13], [14], [15]. During this time, 229 participants contributed towards the first study concerning *sickness* and 471 participants contributed towards the second study concerning *presence* in the 3DTV viewing experience. A further 639 participants contributed towards a third study addressing childrens' responses when watching 3DTV. The results from the third study were very positive, with 71% of the participants saying they "like [3D] very much" compared to just 5% holding a neutral or worse opinion and 73% of participants said they would like to watch 3DTV at home. It was found that 88% of the participants who took part in the first study reported some symptoms of sickness. The second study found that presence was influenced by previously experienced discomfort, whether or not the viewer was standing or sitting and whether or not it was their first 3D viewing experience.

Both the uncontrolled environment and the rapid evaluation methods required for a study conducted in a shopping mall were identified as a limitation of these three studies. Too little information was provided about the 3D content shown to determine what quality it had been designed or measured to have, and we would expect the choice and quality of this content to have a significant impact on the results.

In summary, we conclude that there remains a need to evaluate quality controlled 3D content and its impact on audiences. None of the studies above seek to demonstrate the change in audience perception of 3D films before and after viewing. However, they do indicate that a useful set of concepts to evaluate 3D films are: *viewing experience, naturalness* and *comfort*.

III. Methodology

In this section, we outline the general method used to answer our research questions. This begins with the experimental design in section III-A, followed by the questionnaire design in section III-B. We then give details of the participants recruited for our experiment in section III-C and consider the statistical design of the experiment in section III-D. This section finishes

with a summary of the final general experimental procedure in section III-E. Further details of our methodology, such as the the display and location of each replication, are discussed in later sections.

A. The Experimental Design

As this study is concerned with identifying a change in attitude to 3D films before and after viewing a high quality 3D film, we adopted a one group pre-test post-test quasi-experimental design [16]. This design is simple, effective for identifying change and widely used by researchers. Participants are tested before and after an intervention in order to identify any change in test responses. These response changes are then assumed to be caused by the intervention. In this study the intervention is a 3D film and the tests are questionnaires seeking insight into the participant's attitude towards 3D and awareness of the film's content.

In order to protect the validity of the results the design needs to minimise the effect of any external variables that might impact upon the results. For example, boredom and tiredness, or loss of concentration may occur if the duration of the intervention is too long. The films we presented did not last more than eight minutes, keeping the intervention short. In addition we minimised the effect of other possible external variables by running interventions in a blacked out room and monitoring image brightness and audio volume levels. The test questionnaires run before and after the intervention were kept simple and easy to complete.

We used differentiated replications to investigate how varying key aspects of the intervention affected the audience's responses. Details of each intervention are given in Table I and are discussed below.

We tested responses to two films, *Cosmic Cookery* and *Cosmic Origins*, in order to determine whether the measures we used were stable across similar but different films. Both films were created at Durham University using similar depth budget controls and similar content, but the music, narration and images make them distinctly different films. Details of the original experiment, in which both these films were shown on the 160" large screen projected display, are given in section IV.

We also sought insight into the potential effect from response variance caused by the display technology. In particular we compared results from the large screen projected display (160") with those from using a TV and a small screen projected display (both 50"). Again, we were interested in exploring whether audience responses changed across different viewing platforms. The differentiated replications that used small screens are detailed in sections V and VI.

Finally, we were interested in understanding if the audience responses would vary at locations outside our laboratory in Durham. To do this we ran experiments at the University of York (UK) and overseas at the University of Twente (NL). The display technology used at these locations was the best performing TV sized display from the experiments run in Durham. The details of these differentiated replications are given in section VII.

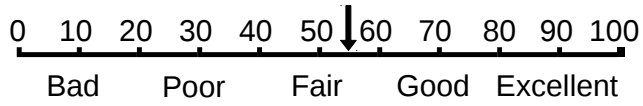

Fig. 1. The response scale used by participants to answer the first five questions in each questionnaire. Participants were asked to indicate their response with an arrow as shown. The print size of this scale was 10 cm long to meet the specifications outlined in [17]

B. The Questionnaires

The preliminary and post-intervention tests were performed using paper questionnaires that began with the same five questions:

1) Please rate your impression of the viewing experience 3D films can provide.
2) Please rate your impression of how well 3D films can convey complex information.
3) Please rate your impression of how comfortable you think viewing 3D films can be.
4) Please rate your impression of how natural the sensation produced by viewing 3D films can be.
5) Please rate your knowledge of how galaxies are made.

Questions 1 and 4 are included with reference to the study by [7] and question 3 with reference to the literature concerning visual discomfort in stereoscopic media [8], [9], [10]. Questions 2 and 5 were added to gather evidence about whether stereoscopic media is a good way of presenting complex, cosmological data. Another question was included in each test, in the preliminary questionnaire this was a closed multiple choice question:

- How would you rate your experience of 3D films? None/Limited/Good/Expert

Whereas in the post-intervention questionnaire it was an open question that included a request for comments:

- Please write any comments or observations you have about 3D films below.

Responses to the first 5 questions were provided by asking participants to draw an arrow on a Likert scale as shown in Figure 1. These scales were designed to meet the recommendations described by the ITU [17]. The indicated values were read off the scales by human eye and recorded in data sheets as integers. The small random error incurred in doing this can be estimated as ± 1.

C. The Participants

The participants were recruited from the academic communities where each round of experimentation was performed. The majority of participants were undergraduate or postgraduate students, though some members of staff also took part. In total, 176 people took part in the study of which 67% were male and 33% female. The ages ranged from 18 to 57, with a median age of 23 and an inter-quartile range from 20 to 26.

All participants were required to give a complete set of responses to the Stereo Titmus Test before their participation.

ID	Location	Display	Film	Coding
D-LP-CC	Durham	160" Projection	Cosmic Cookery	Original SD Resolution
D-LP-CO	Durham	160" Projection	Cosmic Origins	Original HD Resolution
D-TV-CC	Durham	50" TV	Cosmic Cookery	Blu-Ray SD resolution
D-TV-CO-HFR	Durham	50" TV	Cosmic Origins	Blu-Ray Higher frame rate
D-TV-CO-HR	Durham	50" TV	Cosmic Origins	Blu-Ray Higher resolution
D-SP-CC	Durham	50" Projection	Cosmic Cookery	Blu-Ray SD resolution
D-SP-CO-HFR	Durham	50" Projection	Cosmic Origins	Blu-Ray Higher frame rate
Y-SP-CC	York	50" Projection	Cosmic Cookery	Blu-Ray SD resolution
Y-SP-CO-HFR	York	50" Projection	Cosmic Origins	Blu-Ray Higher frame rate
T-SP-CC	Twente	50" Projection	Cosmic Cookery	Blu-Ray SD resolution
T-SP-CO-HFR	Twente	50" Projection	Cosmic Origins	Blu-Ray Higher frame rate

TABLE I

ALL THE INTERVENTIONS EVALUATED ARE SHOWN HERE. THE IDS ARE OF THE FORM LOCATION-DISPLAY-FILM-CODING. WHERE FOR LOCATION: D = DURHAM, Y = YORK, T = TWENTE, FOR DISPLAY TYPE: LP = 160" PROJECTION, TV = 50" TV, SP = 50" PROJECTION, FOR FILM NAME: CO = *Cosmic Origins*, CC = *Cosmic Cookery* AND FOR CODING HR = HIGH RESOLUTION, HFR = HIGH FRAME RATE. THE FIRST GROUP OF TWO INTERVENTIONS WERE OUR FIRST EVALUATIONS ON THE LARGE SCREEN, THE SECOND GROUP OF THREE INTERVENTIONS WERE OUR EVALUATIONS OF THE 50" TV AND THE DIFFERENT POSSIBLE BLUERAY CODINGS FOR CO, THE FINAL GROUP OF SIX INTERVENTIONS WERE THOSE WE SETTLED ON AS SUITABLE FOR EVALUATIONS AT ALL THREE GEOGRAPHIC LOCATIONS USING THE 50" PROJECTION DISPLAY.

Participants who failed to score 100% correct in this test were informed that their results "may not contribute towards the project conclusions" and were invited to choose whether or not to continue their participation, in case their results become of use at a later time. All 56 participants in this situation chose to continue their participation. The study took approximately 30 minutes, for which participants were each paid an honorariam £5, or €5 in the case of our overseas experimentation.

We gathered data until we had at least 15 participants who had passed the screening test in each sample. This sample size of at least 15 is a recommendation from [18] based upon a series of large computational studies [19], [20]. The number of participants who could simultaneously take part in each viewing was dependent upon the screen size of the display technology used. The study was approved by the ethics committee of the School of Engineering and Computing Sciences, Durham University.

D. Statistical Design

Paired Student's t-tests were used to identify whether there was any significant difference between pre and post questionnaire scores across each sample. Student t-tests assume normally distributed samples, so to check this the Shapiro-Wilk test for normality was used. In the case of a sample failing the normality test, the Wilcoxon signed rank test was used instead of the t-test, with the median and inter-quartile range used in place of the mean and standard deviation. In the case of no response difference being identified, two one sided t-tests (TOST) were used to check for equivalence against the null value of zero. All significance testing used an alpha criterion of 0.05 to indicate a "strongly significant result" and 0.10 to indicate a "weakly significant result".

Analysis of Variance (ANOVA) was used to assess the differences between the experiment and replications. Although ANOVA also assumes a normal distribution, it is reputedly insensitive to data normality [21], [22]. We therefore use ANOVA to test between all samples, even where some samples fail Shapiro's test for normality.

E. The Procedure

The procedure required participants to fill out four forms on a clip board. It was decided that the participants should not be allowed to refer to their preliminary responses whilst giving their post-viewing responses. This is because we were seeking a change in attitude towards 3D films, not a self-referenced consideration of the specific film they had viewed. The preliminary questionnaires were therefore collected prior to watching the film and completing the post-viewing questionnaires. The final procedure for each viewing involved the following distinct stages:

1) Welcome participants and outline the procedure to them.
2) Ask them to read and fill out the instructions and consent form.
3) Ask participants to complete the stereo Titmus test by reading and filling out a second form in conjunction with viewing the appropriate images with the appropriate passive glasses.
4) Ask participants to fill out the preliminary questionnaire and then collect all forms in.
5) Hand out the appropriate active glasses for viewing the film and show participants a random dot stereogram to ensure that their glasses are working.
6) Switch lights off and show them the film.
7) Switch the lights on, hand out the post viewing questionnaire and ask them to fill it out.
8) Pay them for their time.

IV. EXPERIMENT: BIG SCREEN PROJECTION

This original experiment used the display technology that we hypothesised was most likely to give positive results — our big screen, low crosstalk, active shutter glasses display system. If an effect was found here for both *Cosmic Origins* and *Cosmic Cookery* we would then have the motivation to consider the other factors of interest.

A. Experimental Setup

The setup for this experiment consisted of:

- Christe Mirage 3D 1080 HD DLP projector

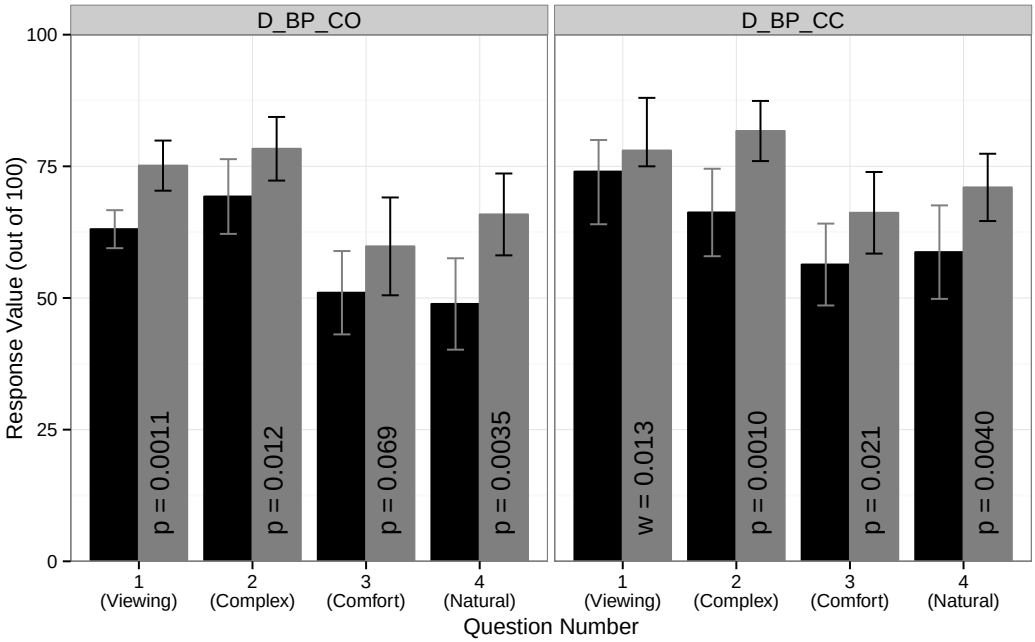

Fig. 2. Showing the results of the original experiment using the big screen projection. Depending upon the result of Shapiro's test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where Shapiro's test failed and ranked statistics are used, the statistical test result is labelled with a *w* instead of a *p*.

- Rear projection screen 3.50 m wide and 1.97 m high
- Virtalis Activeworks 3D Glasses
- JBL EON1500 stereo speaker system

Participants sat in a row centred on the centre of the screen and at a distance such that the central viewer received a 40° viewing angle as recommended by [23]. Five participants completed the experiment at a time. In total 19 participants took part in the *Cosmic Origins* viewings, of which 4 failed the screening test, and 21 participants took part in the *Cosmic Cookery* viewings of which 4 failed the screening test. These participants were recruited primarily through the first year undergraduate engineering course, resulting in an age distribution of 18-32, with a median of 19 and an inter-quartile range of 18-19.

Brightness was measured using a Sekonic L-758 Cine light meter. The receptor was placed behind a "lens" of the active 3D glasses and positioned at approximately the viewing position, with the room darkened as for viewing. A stereo black image pair was shown and the luminance reading through the glasses was found to be too small to detect, meaning that it was less than 0.63 lux. The luminance of a stereo white image pair was found to be 1.3 lux through the glasses.

The maximum volume during the opening few seconds of the narration was measured so that it could be matched in the other experiments. This was done using a decibel meter on a tripod positioned at approximately the central viewer's listening position. The maximum volume for the opening phrase of narration was set at 73.9 dB.

The content was shown at full original-edit quality: *Cosmic Origins* in frame packed 1920x1080 HD with a frame rate of 30 fps and *Cosmic Cookery* in frame packed 1024x768 with a frame rate of 25 fps.

Question 5 was not included in the questionnaires used in this original phase of the study, though we have no reason to believe that this would affect the results in any significant manner.

B. Results

Figure 2 shows summarised results for this experiment including both the *Cosmic Origins* and *Cosmic Cookery* films. For the normally distributed data, a mean preliminary response is indicated by a black bar, whilst a mean post-viewing response is indicated by a grey bar, and the error bars denote the standard deviation.

Responses to each question for each film passed Shapiro's test of normality with a significance criterion of 0.05 in all but one of the eight cases. The post-viewing responses to question 1 (viewing experience) in the *Cosmic Cookery* data yielded a p-value of 0.0107 for Shapiro's test of normality. This is less than our significance criterion, meaning that we need to reject the null hypothesis that the data is normally distributed. We therefore display ranked statistics (median and inter-quartile range) for this question in Figure 2, and used a Wilcoxon Signed Rank Test instead of a Student's t-test to compare preliminary and post-viewing responses. The result of this test is labelled with a *w* in Figure 2 and is smaller than our alpha significance criterion, allowing us to conclude that the response difference is significantly different from zero.

In all cases, except question 3 (comfort) for *Cosmic Origins*, we concluded that the difference between preliminary and

post-viewing responses is strongly significant - the Student's paired t-test or Wilcoxon signed rank test yields a p-value less than our chosen significance criterion of 0.05. The t-test p-value for question 3 (comfort) is 0.069, which is less than 0.1 so we still conclude that it is weakly significant.

The results from this experiment suggest that viewing both *Cosmic Origins* and *Cosmic Cookery* can have a significant effect upon a viewer attitude towards 3D films.

V. Replication 1: Television Display

The effect observed in the original experiment provided motivation for further study seeking significance in other displays. This differentiated replication investigated whether a similar effect is found in Television (TV) displays, which are smaller and make use of very different stereoscopic technologies.

A. Experimental Setup

The following equipment was used:

- Panasonic TXP50ST50B Plasma Active shutter Glasses 3D TV.
- Glasses
- Sony BDP-5780 Blu-ray disc player

The films were played using a 3D Blu-ray disc and player, in order to keep the equipment portable for later use at external sites. As a consequence the films could not be shown in original-edit quality, so we experimented with several encodings to determine the best approach. Using the Sony Vegas software package, we re-encoded the video to the Multiple View Coding (MVC) format, which limited us to a frame rate of 27 fps with full 1080p HD or 60i fps with 720p HD. The conversion from 30 fps to 27 fps was not smooth and caused noticeable jerkiness when viewing. The conversion from 30 fps to 60i fps was smooth, but the loss in resolution was noticeable. We were unsure which encoding would be preferred, so we ran separate viewings for each of 3 different films: 720p HD *Cosmic Origins* with a Higher Frame Rate (HFR) of 60i fps, 27 fps *Cosmic Origins* with a Higher Resolution (HR) of 1080p HD and *Cosmic Cookery*. *Cosmic Cookery* suffered a small loss in resolution as the 1024x768 image was mapped onto a 1280x720 image. The frame rate was 50i fps and the original aspect ratio was maintained, resulting in black space down the left and the right hand sides.

As in the original experiment, participants sat in a row centred on the centre of the screen and at a distance such that the central viewer received a 40° viewing angle. This time, due to the smaller screen size, only three participants could be accomodated in each viewing. The TV was set upon a desk in front of the participants. Twenty participants took part in the *Cosmic Origins* HFR viewings, of which 3 failed the screening test, whilst 17 participants took part in the *Cosmic Origins* HR viewings, of which 2 failed the screening test. Sixteen participants took part in the *Cosmic Cookery* viewings of which 1 failed the screening test. These participants were primarily recruited from the Chemistry, Engineering and Mathematics postgraduate groups, resulting in an age distribution of 19-37, with a median of 24 and an inter-quartile range of 22-26. The gender balance was 53% male to 47% female.

Brightness was measured using the same technique as in section IV-A. The black screen luminance was again less than 0.63 lux whilst the white screen luminance was 1.6 lux. The volume level at the viewer's listening position was matched to the original experiment using a decibel meter.

B. Results

The results of this replication are shown in Figure 3. Three cases failed Shapiro's test for normality, and a Wilcoxon signed rank test was used in place of a Student's t-test to account for this. The preliminary responses to question 1 (viewing experience) in the *Cosmic Origins* HFR data yielded a Shapiro p-value of 0.0359, whilst the post-viewing responses to question 5 (knowledge) in the *Cosmic Origins* HR data yielded a Shapiro p-value of 0.0291. The *Cosmic Cookery* post-viewing responses to question 4 (naturalness) yielded a Shapiro p-value of 0.0129.

The three cases that failed the response difference significance tests are coloured white in Figure 3: question 3 (comfort) for both *Cosmic Origins* films and question 1 (viewing experience) for *Cosmic Cookery*. None of these cases can be considered weakly significant. It is important to note that a failed significance test does not allow us to conclude that no effect exists. Instead, it tells us that there is not enough evidence to conclude whether an effect exists or not. Therefore, we ran equivalence tests designed to determine whether the mean response difference was equal to zero (implying no effect occurred). The significance criterion was taken as 0.05 and a conservative region of equivalence of ± 5 points was chosen, giving an interval width of 10 corresponding to the minor interval on the response scale in Figure 1. No significant result was found. These three cases are therefore null results - they neither support nor oppose the hypothesis that a measurable change in response occurred whilst watching the film. Further discussion is presented in section VIII.

Television displays have yielded a number of significant results suggesting positive changes in response occurred when viewing the films. However, due to the three null results, the effects do do not appear to be as strong as those from the big screen projected display. In section VIII we discuss what might have caused these failed significance tests and how they sit alongside the results from the original experiment.

VI. Replication 2: Small Screen Projection

The TV display gave results with a weaker set of effects than the original experiment. We noticed that our TV display had significantly higher crosstalk than the original projection display – a result of the different imaging technology being used in the display (plasma screen vs DLP projection). This differentiated replication extends the work outlined in the previous section by matching the TV display size using a low crosstalk DLP projection technology as in the original experiment.

A. Experimental Setup

This experiment used the following equipment:

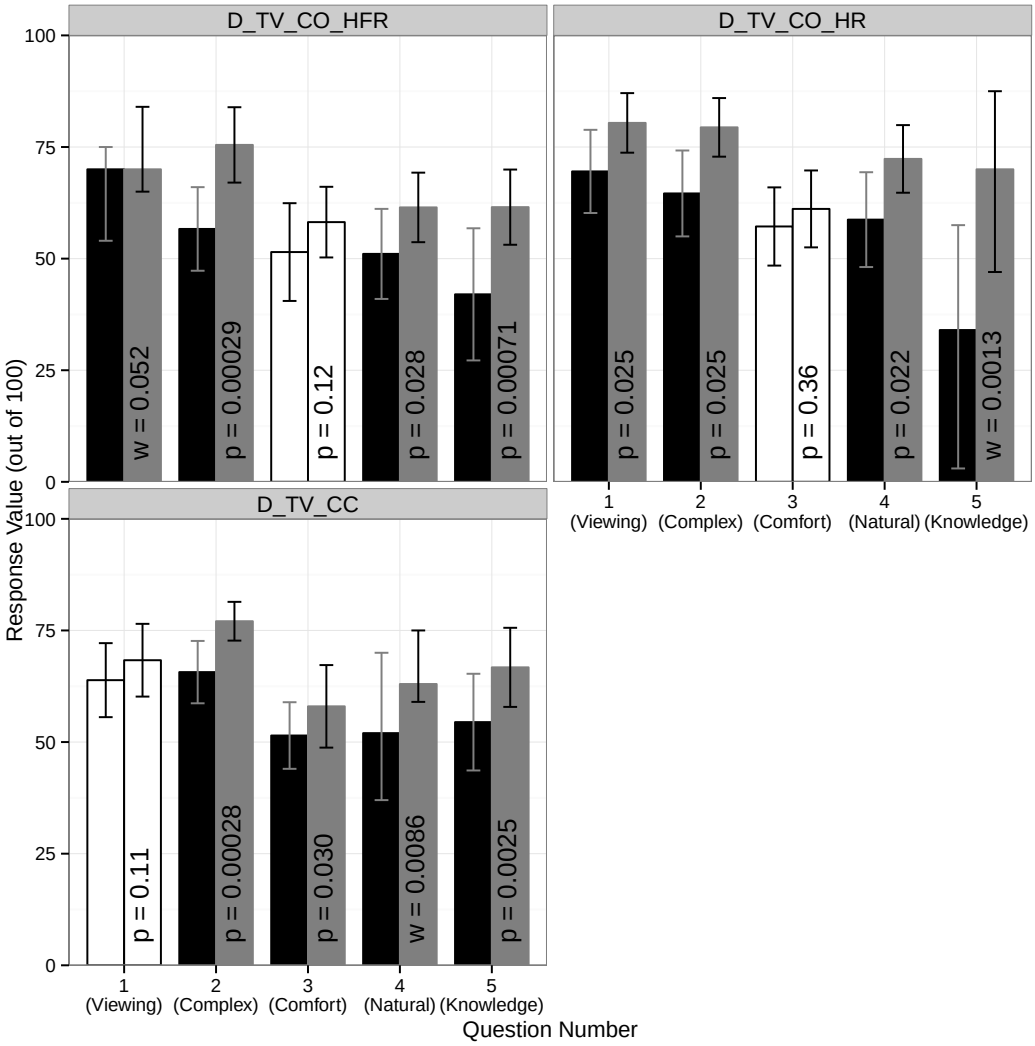

Fig. 3. Showing the results of the differentiated replication using the TV display. Depending upon the result of Shapiro's test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where Shapiro's test failed and ranked statistics are used, the statistical test result is labelled with a *w* instead of a *p*. The white bars indicate questions where the statistical test failed to find a significant difference between preliminary and post viewing responses (the result did not meet our alpha significance critereon of 0.1).

- Optoma HD33-B DLP portable 3D projector
- Optoma ZF2100 glasses and emitter
- Polk-audio Silicon Graphics stereo loudspeaker pair
- Sony BDP-5780 Blu-ray disc player

The films were played using the 3D Blu-ray disc and player, but this time the HR version of *Cosmic Origins* was not shown because in the TV viewings it yielded response differences that were less significant in the majority of cases than those yielded by the HFR version. It also attracted negative comments from the audience in written feedback.

As in the previous replication, three participants at a time sat in a row centred on the centre of the screen and at a distance such that the central viewer received a 40° viewing angle. Twenty-two participants took part in the *Cosmic Origins* viewings, of which three failed the screening test, and 21 participants took part in the *Cosmic Cookery* viewings of which four failed the screening test. These participants were primarily recruited through the second year undergraduate engineering course and a Durham college's postgraduate group, resulting in an age distribution of 19-35, with a median of 21 and an inter-quartile range of 20-23. The gender balance was 61% male to 39% female.

Brightness was measured using the same technique as in section IV-A. The black screen luminance was again less than 0.63 lux whilst the white screen luminance for this screen was notably brighter at 9.3 lux. The volume level at the viewer's listening position was matched to the previous experimentation using a decibel meter.

B. Results

Figure 4 shows the results from the small screen projection viewings. All of the data sets taken using the *Cosmic Origins*

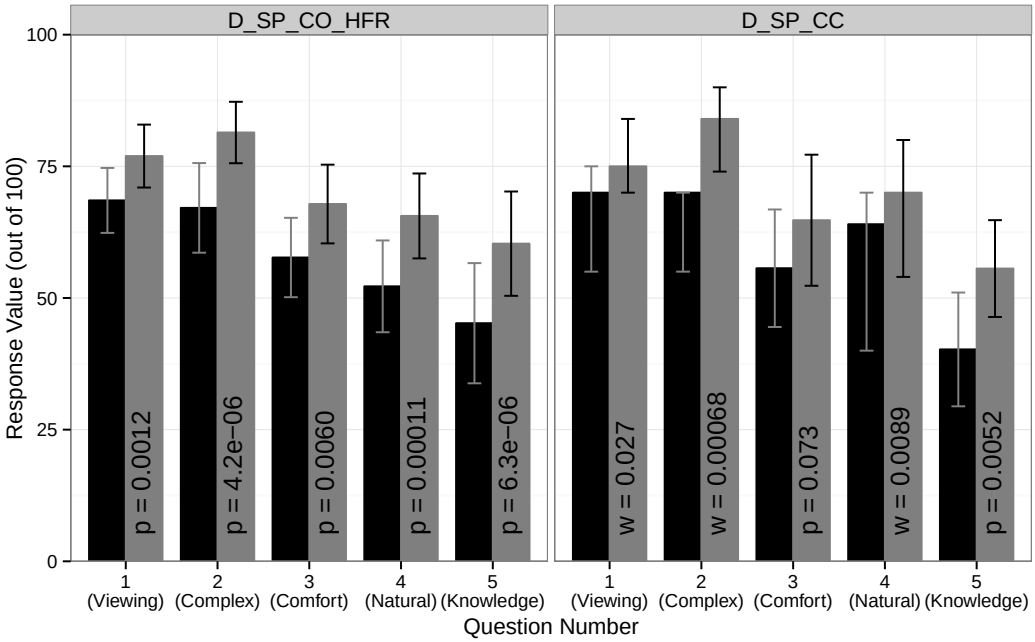

Fig. 4. Showing the results of the differentiated replication using the small screen projection. Depending upon the result of Shapiro's test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where Shapiro's test failed and ranked statistics are used, the statistical test result is labelled with a *w* instead of a *p*.

film passed Shapiro's tests for normality, whilst three questions from the *Cosmic Cookery* data failed the test. Both preliminary and post-viewing responses in question 1 (viewing experience) and question 2 (complex information) failed with respective p-values of 0.0211 and 0.00695 in question 1 and 0.00509 and 0.00242 in question 2. Shapiro's test also failed in question 4 (naturalness) with preliminary responses yielding a p-value of 0.023.

The only significance test to yield a result that was not strongly significant is question 3 (comfort) for the *Cosmic Cookery* data. The Student's t-test gives a p-value of 0.0727, which indicates a weakly significant effect. These results are therefore similar to the big screen results, despite the significant amount of compression applied to the films so that they could be played from a Blu-ray disc. The data also shows that a more significant effect occurred than when watching the films on the TV display. As a result we chose the small screen projected display to evaluate response differences outside our laboratory at Durham.

VII. Replications 3 & 4: York and Twente

We next sought to demonstrate that our results are repeatable beyond our own laboratory and the academic community where the films were created. This was done by taking the best performing portable display - the small screen projection - first to another site in the UK, and then further afield to an international site in the Netherlands.

A. Experimental Setup

This differentiated replication used the equipment outlined in section VI-A. The equipment was taken to rooms in the University of York and Twente University and set up in the same way. Using the technique outlined in section IV-A, the black screen and white screen luminance at both sites were measured to be less than 0.93 lux and 9.3 lux respectively. The volume level at the viewer's listening position was matched to the previous experimentation using a decibel meter.

At the University of York participants were recruited from the undergraduate and postgraduate courses run in the Department of Theatre, Film and Television and the Department of Computer Science. Some members of staff also took part. Eighteen participants undertook *Cosmic Origins* HFR viewings, of which 1 failed the screening test, and 24 participants took part in the *Cosmic Cookery* viewings of which 5 failed the screening test. Ages were distributed between 18-57, with an interquartile range of 19-26.75 and a median of 21. The gender balance was 71% male to 29% female.

The experimentation at Twente was run during the summer holidays, so participants could not be recruited from the undergraduate body. Instead they were sourced primarily using postgraduate and staff mailing lists. Twenty-one participants took part in the *Cosmic Origins* HFR viewings, of which 4 failed the screening test, and 22 participants took part in the *Cosmic Cookery* viewings, of which 5 failed the screening test. Ages were distributed between 22-38, with an interquartile range of 24-28 and a median of 26. The gender balance was 80% male to 20% female.

B. Results

The results for the experimentation undertaken at the University of York are shown in the first two graphs of Figure 5.

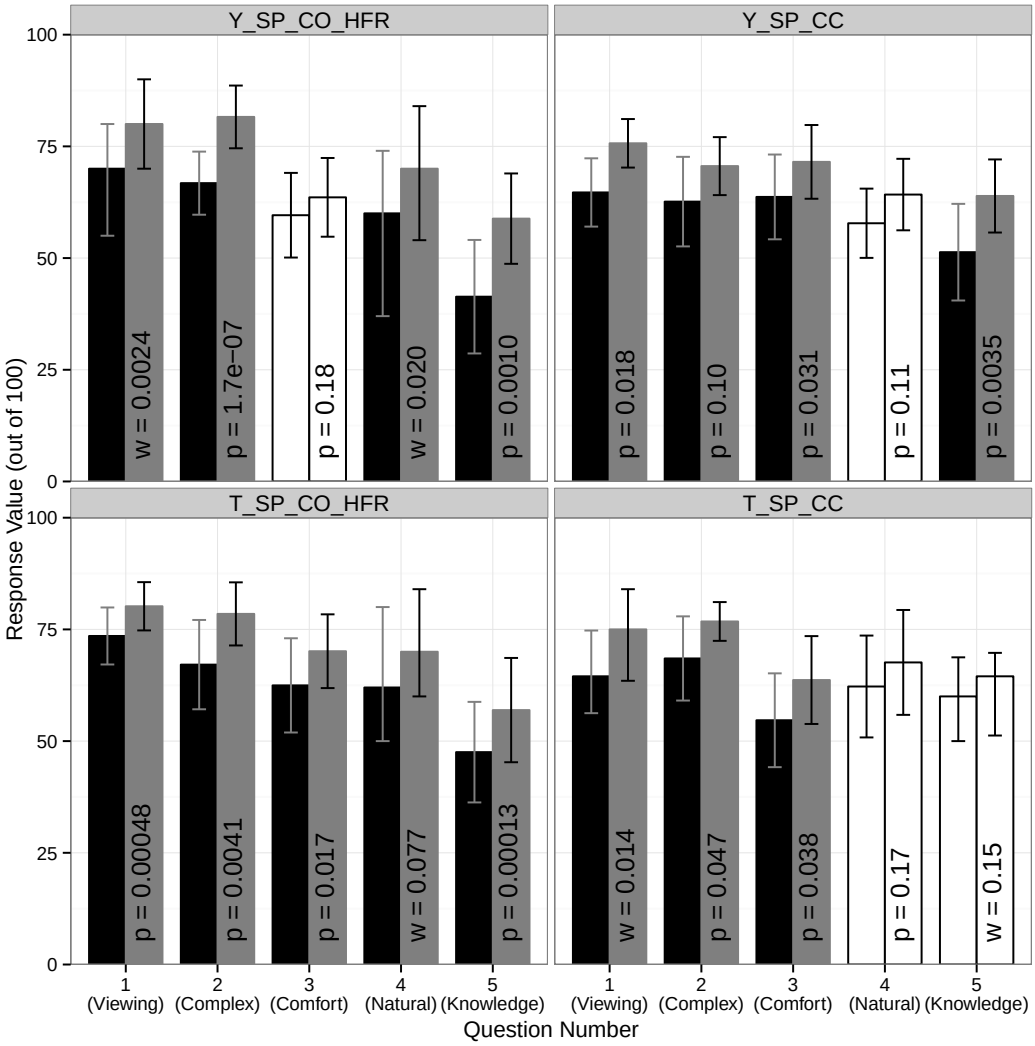

Fig. 5. Showing the results of experiment 4 using the small screen projection at sites in York (UK) and Twente (The Netherlands). Depending upon the result of Shapiro's test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where Shapiro's test failed and ranked statistics are used, the statistical test result is labelled with a *w* instead of a *p*. The white bars indicate questions where the statistical test failed to find a significant difference between preliminary and post viewing responses (the result did not break our alpha significance criterion of 0.1).

Only the post-viewing responses to question 1 (viewing experience) and the preliminary responses to question 4 (naturalness) failed Shapiro's test for normality with p-values 0.0308 and 0.0266 respectively. Response differences to Question 3 (comfort) in the *Cosmic Origins* data, and Questions 2 (complex information) and 4 (naturalness) in the *Cosmic Cookery* data, yielded failed significance tests. Equivalence tests show that these mean response differences are not equal to zero, so we conclude that they are null results (like those discussed in section V-B).

The Twente results are shown in the latter two graphs of Figure 5. The post viewing responses to Question 4 (naturalness) was the only data set in the *Cosmic Origins* data to fail Shapiro's test for normality with a p-value of 0.0296. The post-viewing responses to Question 1 (viewing experience) and the preliminary responses to Question 5 (knowledge) in the *Cosmic Cookery* data failed Shapiro's test for normality with p-values of 0.0152 and 0.0399 respectively. Questions 4 (naturalness) and 5 (knowledge) from the *Cosmic Cookery* data failed to pass the significance tests.

VIII. Discussion

We begin by reviewing the individual cases where our significance testing was successful (section VIII-A), before turning to speculate on those cases where it was not (section VIII-B). We then use ANOVA to identify differences within the data (section VIII-C), which is followed by an analysis of the combined data taken from all our experimentation (section VIII-D). This section concludes by discussing threats to the validity of our results (section VIII-E).

ID	Question 1		Question 2		Question 3		Question 4		Question 5	
	Test	p-Value	Test	p-Value	Test	p-Value	Test	p-Value	Test	p-Value
D-BP-CO	t	0.0011	t	0.012	t	0.069	t	0.0035	-	-
D-BP-CC	w	0.013	t	0.0010	t	0.021	t	0.0040	-	-
D-TV-CO-HFR	w	0.053	t	2.9E-4	**t**	**0.12**	t	0.028	t	7.1E-4
D-TV-CO-HR	t	0.025	t	0.025	**t**	**0.36**	t	0.022	w	0.0013
D-TV-CC	**t**	**0.11**	t	2.8E-4	t	0.030	w	0.0086	t	0.0026
D-SP-CO-HFR	t	0.0012	t	4.2E-6	t	0.0060	t	1.1E-4	t	6.3E-6
D-SP-CC	w	0.027	w	6.8E-4	t	0.073	w	0.0089	t	0.0052
Y-SP-CO-HFR	w	0.0024	t	1.7E-7	**t**	**0.18**	w	0.020	t	0.0010
Y-SP-CC	t	0.018	t	0.10	t	0.031	**t**	**0.11**	t	0.0035
T-SP-CO-HFR	t	0.00048	t	0.0041	t	0.017	w	0.077	t	0.00013
T-SP-CC	w	0.014	t	0.047	t	0.039	**t**	**0.17**	**w**	**0.15**

TABLE II
SHOWING ALL THE P-VALUES FROM SIGNIFICANCE TESTS DETERMINING WHETHER WE CAN REJECT THE NULL HYPOTHESIS THAT THERE IS NO CHANGE BETWEEN PRELIMINARY AND POST-VIEWING RESPONSES. THE ID SYMBOL IS BROKEN INTO THREE PARTS. THE FIRST LETTER INDICATES THE SITE: D FOR DURHAM, Y FOR YORK AND T FOR TWENTE. THE SECOND TWO LETTERS INDICATE THE DISPLAY: BP FOR BIG PROJECTOR, TV FOR TELEVISION AND SP FOR SMALL PROJECTOR. THE FINAL SET OF LETTERS INDICATE THE FILM: CO FOR *Cosmic Origins* AND CC FOR *Cosmic Cookery*. AS MULTIPLE VERSIONS OF *Cosmic Origins* HAVE BEEN USED A FURTHER TWO LETTERS ARE USED: HR CORRESPONDS TO THE HIGHER RESOLUTION VERSION AND HFR CORRESPONDS TO THE HIGHER FRAME RATE VERSION.

A. Significance Test Successes

The p-values from all significance t-tests are shown in Table II. They show that the results are overwhelmingly positive, with the majority (79%) of significance tests yielding a "strongly significant" result. questions 1 (viewing experience), 2 (complex information) and 5 (knowledge) performed particularly well, with only one significance failure for each.

Question 2 was the strongest performing question in this study, with a mean response difference of 15.17 and all cases proving at least weakly significant. Furthermore, there was only one experiment in which the significance test for Question 2 did not prove strongly significant. These strong results are supported by the comments of 23 participants that suggest 3D is particularly suitable for conveying complex spatial information. For instance, *"Watching 3D films may improve and enhance understanding, particularly on complex topics which need 3D graphics to emphasise a point."* It seems that the binocular cue can improve the processing of complex visual information.

The results from Question 5 (knowledge) also performed well, with only one significance failure and an average response difference of 15.09. A small number of comments contrast with these strong numeric results by arguing that the visuals distracted them from the film's narration. One such comment said, "Sometimes the 3D effects can distract from the narration as I found I was too focused on the visuals." It would be interesting to undertake further study assessing the impact of 3D visuals upon processing audio-visual information.

Twelve participants stated in their comments that the purpose of 3D in films needs further consideration. One such individual said 3D effects "have tended to be seen as a gimmick rather than a form of visual expression. If we can move away from the sensationalist "theme ride" nature of current 3D viewing [it] could be very effective." This suggests, then, that the 3D effect should offer added value in the content. Such added value may be found in complex visual information, of which the content in *Cosmic Origins* and *Cosmic Cookery* is an example.

B. Significance Test Failures

The seven results that failed to prove even weakly significant are shown in bold in Table II. Here, we speculate on why these cases failed to show significance.

Three of the null results occurred when viewing the films on the TV display. When analysing the comments we found that 19% of participants who took part in the TV viewings actively complained about crosstalk. Whereas only one comment from the rest of the experimentation could potentially be connected to crosstalk: "Images are still split into two when they come further away from the screen". Crosstalk is a negative factor associated with the 3D displays that may possibly explain these three failed significance tests [24].

Question 3 (comfort) yielded the weakest set of results (3 out of 11 cases failed to prove even weakly significant). It seems that discomfort can still be a problem even when when viewing 3D films with quality controlled depth. Analysing the comments can perhaps offer some further insight into this matter. Whilst 23 participants did complain about discomfort/ache/tiredness specifically in the eyes, almost the same number (22) complained about discomfort due to wearing glasses — a factor that cannot be influenced by high quality content. There were a number of comments concerning comfort that were very favourable, such as, "The film seen today was noticeably more comfortable to watch than normal 3D films." A few people acknowledged improved comfort whilst questioning whether this would hold for longer time periods, such as "Obviously, I have just watched a brilliant 3D film and feel comfortable. I just wonder whether the technique of the short film can be successfully applied to other long films." The short length of each film is a limitation of this study since visual comfort can degrade over viewing time [8], [10].

All significance failures, except those in question 3 (comfort), occur in *Cosmic Cookery* viewings. It is hard to see why *Cosmic Cookery* performs so erratically, with failures

in every question except number 3 (comfort). It seems most likely these failed significance tests are the result of the small sample size limiting the statistical power. When designing this experiment we sought to achieve the commonly accepted value for statistical power of 80%. For a sample size of 15, with standard deviation and effect size set at 10 scale units the statistical power is actually found to be 85%. However, this still suggests that we should fail to correctly reject the null hypothesis in 15% of the t-tests. In actual fact our t-tests have failed in 6 of 43 cases, which is equivalent to 14% of the tests. If we were to repeat the experiment, we would consider using samples of approximately double the size, to attain 98.5% power. Whilst the statistical power may explain our failed t-tests, it does not threaten the validity of conclusions drawn from successful tests.

C. Looking for Differences with ANOVA

Analysis of Variance (ANOVA) performed across all 11 studies for questions 1-4 yielded no significant differences between studies. Table III shows the F-values and the probabilities associated with these ANOVA. The only question with any significant difference between studies is question 5 (knowledge). The participants for this study have been recruited from selected academic communities. One could expect differences to occur in the learning of content information, and thus response differences to question 5, based upon the academic discipline (i.e. Maths students may be more interested in, and better prepared to learn about, galaxy formation than Anthropology students). As the recruiting of participants often involved targeting specific groups of academics, each sample of participants did not represent a random selection across academic disciplines. This could explain the variance observed in question 5.

The failed ANOVA tells us that there is not enough evidence to conclude that the contributing samples are taken from different distributions. Therefore, analysis of the combined data (from all rounds of the experimentation) may be of interest. For each question that failed the ANOVA, Table III also includes the details of t-tests that have been performed using combined data. Every test passes, including the erratic question 3 concerning comfort. We can also conclude from these ANOVA that the results are repeatable for different films, sites and display technologies.

D. Analysing Combined Data

Figure 6 shows the results of combining data from all rounds of experimentation. In total, 186 participants contributed to this combined data set. Student t-tests were run on each film's combined data to establish whether there were significant differences between preliminary and post-viewing responses. All tests yielded a strongly significant results.

For each of the first four questions the combined data was split by gender and the means and standard deviations of each gender's responses to each question calculated. Independent two sample t-tests for samples with unequal sizes and variance were then used to determine if the mean responses differed significantly with gender. No significance was found, suggesting that gender is not an influencing factor upon the observed change in attitude towards 3D films.

E. Threats to the validity of our results

The steps we have taken to minimise threats to the *construct validity* of our results have already been discussed in section III-A. By using short films, simple questionnaires and controlling certain aspects of the environment, we have removed a number of factors that literature suggests may threaten the existence of a causal relationship between our intervention (the 3D film viewing) and the differences in the test results (the questionnaire response differences).

Unfortunately the presence of significant threats to the *internal validity* of our results cannot be ruled out, because we were unable to find a suitable intervention for a control study. There is no accepted definition of a "normal" 3D film for us to test our "high quality" 3D films against. A pre-test post-test quasi-experimental design is often used when no control is available, as the preliminary responses act in a similar manner to a control study for the post-intervention results to be compared against. The preliminary responses rule out any bias caused by prior experience of 3D film quality. Consequently, if we can trust that participants answered our questions honestly and appropriately, and were not led to do otherwise by some aspect of the experiment's execution other than the intervention, then we can trust the validity of our results.

One outstanding threat to the internal validity of our results

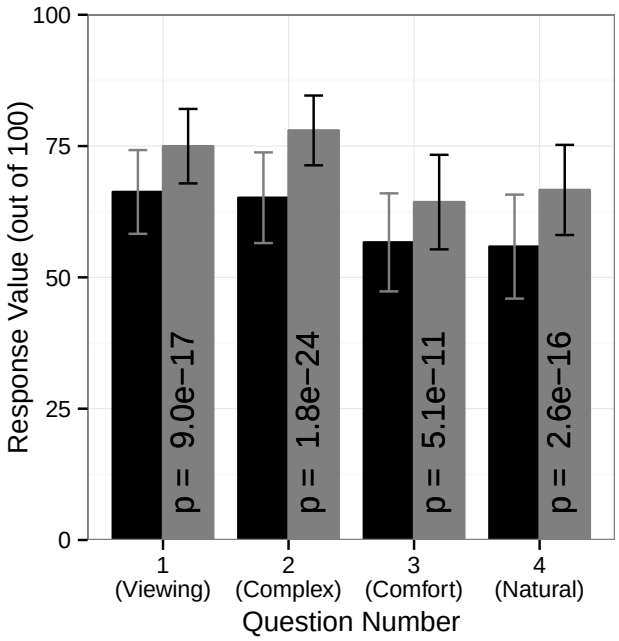

Fig. 6. Showing the results of combining all our data from 186 participants who passed the screening test. The black bars indicate the mean preliminary response, whilst the grey bars indicate the mean post-viewing response and the error bars denote the standard deviation in responses. The p-value (labelled *p*) of a paired Student's t-test is also shown for each question.

Question	ANOVA F-Value	ANOVA Pr(F)	Combined Data (n=186) Mean	Combined Data (n=186) Std. Dev.	Combined Data (n=186) p-value
1 Viewing Experience	0.53	0.87	8.715	12.96	9.0e-17
2 Complex Information	0.85	0.58	12.82	14.75	1.8e-24
3 Comfort	0.36	0.96	7.672	14.99	5.1e-11
4 Naturalness	0.86	0.57	10.79	16.34	2.6e-16
5 Knowledge	3.6	8.2E-4	-	-	-

TABLE III
SHOWING THE RESULTS OF ANOVA SEEKING ANY DIFFERENCES BETWEEN THE ORIGINAL EXPERIMENT AND DIFFERENTIATED REPLICATIONS FOR EACH QUESTION. WHERE THE ANOVA FAILED TO FIND ANY DIFFERENCES, DETAILS OF T-TESTS USING THE COMBINED DATA ACROSS ALL EXPERIMENTATION ARE GIVEN. THESE T-TESTS AGAIN USE THE NULL HYPOTHESIS THAT THE MEAN RESPONSE DIFFERENCE IS ZERO. THE ALPHA SIGNIFICANCE CRITERION WAS 0.05, SO ONLY QUESTION 5 YIELDED A SIGNIFICANT ANOVA RESULT, WHILST ALL OF THE COMBINED DATA T-TESTS PROVED SIGNIFICANT.

is the brightness variation between the three different screens. However, the brightness for each screen showing a stereo white image pair remained within the normal range of indoor office illumination [25]. We therefore wouldn't expect the brightness differences to affect stereo depth perception. Furthermore a one-way ANOVA across all the experiments does not show any significant difference in the results for question 1-4, meaning we have no evidence that the screen technology, incorporating the brightness, affected the audiences' subjective impressions of our films.

This study is made up of differentiated replications of the same experiment, using different participants, films, displays and sites to gain a wider understanding of the scope of our results. Despite this, it is important for us to acknowledge that there are bounds to the scope, which pose threats to the *external validity* of our results. We can conclude very little concerning the bounds of the scope, so researchers should be careful about assuming our results hold in scenarios with notably different characteristics. For instance, our participant samples were not truly random, as they were sourced from academic communities of students and researchers, so would typically be dominated by a particular academic discipline and a particular age group. Therefore, our results may not hold for audiences with a significantly different demographic, such as those made up of children or the elderly.

IX. CONCLUSIONS

In this study we have shown 3D films with quality controlled binocular depth to groups of participants. Before and after watching the film we asked the participants to fill out a questionnaire. Both questionnaires asked the same questions concerning their attitude towards 3D films. Responses were given on a 0-100 point scale, where a greater number indicated a more positive response. This paper reports an original experiment and four differentiated replications, across which we varied the display, film and site used. The original experiment investigated reactions to a large screen projected display in our Durham based laboratory This was followed by replications using a TV display and a small (TV-sized) projected display. The small projected display was then taken off-site to the University of York (UK) and the University of Twente (The Netherlands). The films that we used were created by a collaboration of physicists and computer scientists at Durham University and were entitled *Cosmic Origins* and *Cosmic Cookery*. Between 15 and 19 participants who had been successfully screened for stereo vision took part in each viewing. The difference between their preliminary and post viewing questionnaires were tested against the null hypothesis that they would be equal to zero. Paired Student t-tests or Wilcoxon signed rank Tests were used as appropriate to determine the confidence with which we could reject this null hypothesis, and say that a response change had occurred across the audience. ANOVA were used to look for differences in means between the original experiment and replications. The statistical results were discussed alongside comments left by participants at the end of the post-viewing questionnaire.

In answer to our first research question (section I), we have seen that high quality 3D films using quality-controlled binocular cues can create a measurable positive change in an audience's attitude towards 3D films. This change was observed in response to all of the following questions:

1) Please rate your impression of the viewing experience 3D films can provide.
2) Please rate your impression of how well 3D films can convey complex information.
3) Please rate your impression of how comfortable you think viewing 3D films can be.
4) Please rate your impression of how natural the sensation produced by viewing 3D films can be.
5) Please rate your knowledge of how galaxies are made.

Use of ANOVA failed to find any differences between the experiment and replications in response changes to each of the first four questions. It is possible, then, that each data sample comes from the same distribution. Paired Student's t-tests between preliminary and post-viewing responses across the combined data gave strongly significant results for the first four questions. This therefore indicates that the positive changes in attitude towards 3D films that have been observed in Questions 1-4 are repeatable at national and international sites, as well as for different display technologies and quality-controlled film content. Significant differences in response changes were found between the experiment and replications for Question 5 (knowledge). We have speculated on whether this is due to participants being recruited through specific academic disciplines.

This study motivates research concerning high quality 3D content creation by showing that such content elicits measurable, repeatable changes in audience attitude towards 3D. Furthermore, these attitude changes remain significant for different displays, sites and high quality content. Our research

therefore concludes that the current popular attitude towards 3D may be significantly improved by the wider distribution of high quality content, created with algorithms such as outlined by [1] and [2].

ACKNOWLEDGMENT

The authors would like to thank the research office at Durham University for funding this study and in addition the University of York and the VR Laboratory at Twente University in the Netherlands for supporting the use of their facilities. In addition we thank the ICC at Durham University for their ongoing collaboration in making the 3D films used in this study. Finally, we would also like to thank the anonymous reviewer for their helpful comments.

REFERENCES

[1] G. Jones, D. Lee, N. Holliman, and D. Ezra, "Controlling perceived depth in stereoscopic images," *Stereoscopic Displays and Virtual Reality Systems*, vol. 4297, pp. 42–53, 2001.

[2] N. Holliman, "Mapping perceived depth to regions of interest in stereoscopic images," in *Proceedings of Stereoscopic Displays and Virtual Reality Systems XI - SPIE Volume 5291*, 2004.

[3] R. M. Lindsay and A. S. C. Ehrenberg, "The design of replicated studies," *The American Statistician*, vol. 47, no. 3, pp. pp. 217–228, 1993.

[4] M. Hassenzahl and N. Tractinsky, "User experience–a research agenda," *Behaviour and Information Technology*, vol. 25, no. 2, pp. 91–97, 2006.

[5] N. Holliman, C. Baugh, C. Frenk, A. Jenkins, B. Froner, D. Hassaine, J. Helly, N. Metcalfe, and T. Okamoto, "Cosmic cookery: making a stereoscopic 3D animated movie," in *Proceedings of Stereoscopic Displays and Virtual Reality Systems XIII - SPIE Volume 6055*, 2006.

[6] N. Holliman, "Cosmic origins: experiences making a stereoscopic scientific movie," in *Proceedings of Stereoscopic Displays and Applications XXI - SPIE Volume 7237*, 2010.

[7] P. J. Seuntins, I. E. Heynderickx, W. A. IJsselsteijn, P. M. J. van den Avoort, J. Berentsen, I. J. Dalm, M. T. Lambooij, and W. Oosting, "Viewing experience and naturalness of 3D images," in *Proceedings of Three-Dimensional TV, Video, and Display IV - SPIE Volume 6016*, vol. 6016, 2005, pp. 601 605–601 605–7. [Online]. Available: http://dx.doi.org/10.1117/12.627515

[8] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, "Visual discomfort and visual fatigue of stereoscopic displays: a review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 30 201–1, 2009.

[9] K. Ukai and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays*, vol. 29, no. 2, pp. 106 – 116, 2008.

[10] Y. Nojiri, H. Yamanoue, A. Hanazato, M. Emoto, and F. Okano, "Visual comfort/discomfort and visual fatigue caused by stereoscopic hdtv viewing," in *Proceedings of Stereoscopic Displays and Virtual Reality Systems XI - SPIE Volume 5291*, 2004, pp. 303–313.

[11] L. Lipton, *The foundations of stereoscopic cinema*. Van Nostrand Reinhold Company, 1982.

[12] M. Pölönen, M. Salmimaa, J. Takatalo, and J. Häkkinen, "Subjective experiences of watching stereoscopic avatar and U2 3D in a cinema," *Journal of Electronic Imaging*, vol. 21, no. 1, pp. 011 006–1–011 006–8, 2012.

[13] M. Obrist, D. Wurhofer, F. Förster, T. Meneweger, T. Grill, D. Wilfinger, and M. Tscheligi, "Perceived 3DTV viewing in the public: insights from a three-day field evaluation study," in *Proceedings of the 9th international interactive conference on Interactive television*, ser. EuroITV '11. New York, NY, USA: ACM, 2011, pp. 167–176.

[14] M. Obrist, D. Wurhofer, M. Gärtner, F. Förster, and M. Tscheligi, "Exploring children's 3DTV experience," in *Proceedings of the 10th European conference on Interactive tv and video*, ser. EuroiTV '12. New York, NY, USA: ACM, 2012, pp. 125–134.

[15] M. Obrist, D. Wurhofer, T. Meneweger, T. Grill, and M. Tscheligi, "Viewing experience of 3DTV: An exploration of the feeling of sickness and presence in a shopping mall," *Entertainment Computing*, vol. 4, pp. 71–81, 2013.

[16] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, 2002.

[17] ITU-R Recommendation BT.500-12, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., 2009.

[18] D. S. Moore, *The Basic Practise of Statistics*, 2nd ed. Macmillan Education Australia, 1995.

[19] E. S. Pearson and N. W. Please, "Relation between the shape of population distribution and the robustness of four simple test statistics," *Biometrika*, vol. 62, no. 2, pp. 223–241, 1975.

[20] H. O. Posten, "The robustness of the one-sample t-test over the pearson system," *Journal of Statistical Computation and Simulation*, vol. 9, no. 2, pp. 133–149, 1979.

[21] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance," *Review of Educational Research*, vol. 42, no. 3, pp. pp. 237–288, 1972.

[22] L. M. Lix, J. C. Keselman, and H. J. Keselman, "Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test," *Review of Educational Research*, vol. 66, no. 4, pp. 579–619, 1996.

[23] THX, "HDTV set up," http://www.thx.com/consumer/home-entertainment/home-theater/hdtv-set-up/, June 2013, (Accessed on this date).

[24] S. Pala, R. Stevens, and P. Surman, "Optical cross-talk and visual comfort of a stereoscopic display used in a real-time application," in *Electronic Imaging 2007*. International Society for Optics and Photonics, 2007, pp. 649 011–649 011.

[25] G. D. Ander, *Daylighting performance and design*. John Wiley & Sons, 2003.

Jonathan Berry received his BSc joint honours degree in physics and computer science from Durham University (UK), where he now studies for a PhD in the department of Engineering and Computing Sciences. His PhD research, supervised by Nick Holliman and David Budgen, concerns the development of quality-controlled content for stereoscopic displays and spatial sound systems, with a particular focus on audio-visual interactions that come to light when combining the two technologies.

David Budgen received the BSc (Hons.) degree in physics and the PhD degree in theoretical physics from Durham University (UK). He worked as a research scientist for the Admiralty and then held academic positions at Stirling University and Keele University before moving to his present post as a professor of software engineering at Durham University in 2005. His research interests include software design, design environments, empirical (evidence-based) software engineering and healthcare computing.

Nick Holliman is a professor of interactive media at the University of York (UK). He researches the science and engineering of interactive media including the fundamental challenges of stereoscopic 3D visualisation. This includes working with psychologists to understand how the human visual system processes binocular information, developing novel computational algorithms for the control of binocular image disparity and demonstrating how these algorithms work in practice in software tools and award winning 3D visualisations of cosmology data.